国立国語研究所学術情報リポジトリ

An Examination of Stepwise Specialization of Morphological Analysis Dictionary “UniDic” for Efficient Word Annotation on “Corpus of Historical Japanese” : The Case of Chikamatsu Corpus

メタデータ	言語: jpn 出版者: 公開日: 2018-03-20 キーワード (Ja): キーワード (En): 作成者: 岡, 照晃, OKA, Teruaki メールアドレス: 所属:
URL	https://doi.org/10.15084/00001476

『日本語歴史コーパス』短単位アノテーション作業効率化に向けた形態素解析用辞書『UniDic』の段階的特殊化の検討 –近松コーパスを例として–

岡 照晃 (国立国語研究所コーパス開発センター) *

An Examination of Stepwise Specialization of Morphological Analysis Dictionary "UniDic" for Efficient Word Annotation on "Corpus of Historical Japanese" –The Case of Chikamatsu Corpus–

Teruaki Oka (National Institute for Japanese Language and Linguistics)

要旨

本論文では，現在，国語研の通時コーパス構築プロジェクトで整備中の近世前期の上方資料である『近松門左衛門 世話物浄瑠璃』への短単位形態論情報アノテーションの効率化を目的に，形態素解析器 MeCab の追加学習機能を使い，既存の『洒落本』用の短単位解析用辞書から段階的に，近松専用短単位解析用辞書を作成する方法について述べる．具体的には，まず比較的時代の近い洒落本解析用辞書を，上方の洒落本コーパスのみで上方の洒落本解析用辞書にアダプテーションする．次に作成した上方の洒落本解析用辞書を，同じく上方の資料である近松コーパスで近松資料解析用辞書にさらにアダプテーションする．本手法により，従来手法よりも高い精度（語彙素認定 F1 値，地の文：86.85 → 89.60，会話文：85.07 → 88.82）で，近松資料を解析できることを確認した．また本論文で作成した短単位解析用辞書を使った近松資料のコーパス化作業が現在，進行中である．

1. はじめに

国立国語研究所では現在，日本語の通時コーパス『日本語歴史コーパス』(近藤泰弘 2012)（以下，**CHJ**）の構築を進めている．CHJ の特徴として，国語研の規定する言語単位**短単位** (伝康晴ほか 2007, 近藤泰弘 2015) での形態論情報（e.g., 品詞，活用，発音，語種…）がアノテーションされていることがある．この形態論情報の人手アノテーション効率化のため，CHJ 構築では各時代専用に形態素解析器『MeCab』(Kudo et al. 2004) 用の『解析用辞書 UniDic』(小木曽智信ほか 2013, 小木曽智信ほか 2014)（以下，単に**辞書**という場合，この**解析用辞書**を指す）のコストを学習し，MeCab による自動解析の後，解析結果を人手修正する，という作業方針を採っている．そのため自動形態素解析の結果の精度は人手作業の負担と直結し

* teruaki-oka {at} ninjal.ac.jp

ている．もし自動解析結果の精度が高ければ，人手修正が必要な箇所も少なく，形態論情報アノテーションにかかる人的・時間的負担も少ない．しかし反対に自動解析結果の精度が低いと，（人手で 0 からアノテーションするよりも格段に少労力だが）作業にかかる負担は大きい．

現状において，辞書は文献 (鴻野知暁ほか 2014) に基づき，各時代ごとに 1〜2 個（文語，口語）という粒度で作成されている．これは MeCab の内部モデルである CRF (Lafferty et al. 2001) の学習用コーパスをまとまった量で確保し，解析結果の精度を高めるためと，整備対象となる多様な資料に対して，できるだけ汎用的に使用でき，再利用可能な辞書を目指したためである．この汎用化は，外部の研究者に作成した辞書を公開することを視野に入れた考慮でもある．

これに対し本論文では，CHJ 構築現場での形態論情報アノテーション作業のさらなる効率化を第一目的に，より解析対象の資料に適合した辞書の構築について述べる．汎用的な辞書は多様な資料に使用できる反面，全体で見た時の精度は高くとも，各ドメインに特徴的な箇所の多くで，解析エラーが生じやすくなる．またドメインが一様でないコーパスを混用して学習するため，各ドメイン間でサイズが異なる場合は，その影響を受けて解析も偏りやすい(1)．実際，文献 (市村太郎 2014) では，近世の資料の地の文と会話文の文体・文法体系の差を指摘し，文献 (市村太郎ほか 2016) において，それ以前では学習時に混用されていた地の文と会話文を分け，それぞれに専用の辞書を学習することで解析結果の精度向上を報告している．

各資料に特化した辞書を作成する際の問題は，汎用的な辞書を作成する場合よりも学習用コーパスが絞り込まれてしまうことがある．そこで提案手法では，これまで通りの汎用辞書（のモデル）を始点として，段階的に学習用コーパスを絞り，MeCab の**追加学習**(2)機能によって，汎用辞書を段階的に特殊化していく方針を採る．この手法の利点として，① 整備中の資料であっても少量の人手アノテーションが終われば，逐次，辞書の学習に追加していけること，② 各時点での辞書を別の資料用辞書の学習の始点として再利用できること，③ これまで通り汎用の解析用辞書は作成していくため，国語研内部だけの特殊化した辞書と同時に，外部の研究者に有用な公開用辞書も同時に作成できることがある．

本論文では，CHJ に含まれる近世の後期の資料『洒落本』（以下，**CHJ 洒落本**）および，現在整備を進めている同じく近世の上期上方資料『近松門左衛門 世話物浄瑠璃』（以下，**CHJ 近松**）を対象に，洒落本用辞書の学習から開始し，江戸・上方各専用の辞書を作成，さらに上方専用の辞書から近松専用の辞書を作成する流れを説明する．

2. 近世期資料『洒落本』の自動形態素解析の現状と問題

CHJ では，古代語〜近代語についての代表的な資料を集め，日本語の通時的な研究に利用可能なコーパスとして順次公開を進めている．このうち『**CHJ 江戸時代編**』としてコーパス化の対象となっているのが，近世後期の洒落本と人情本，そして近世前期の上方資料であ

(1) 注） 学習時（or 前），何かしらの正規化により，このサイズのアンバランスを解消する場合は別．

(2) MeCab の公式ページ: `http://taku910.github.io/mecab/learn.html` (2017/2/4 現在) では「再学習」と記載されている機能であるが，「辞書を 0 から学習し直す」という意味と混同しやすく，また同ページに「再学習とは，現在の学習済みモデルと少量の追加学習データからモデルを再構築する仕組みです」とあることから，本論文ではこれを「追加学習」と呼称する．

表 1　本稿で使用した CHJ 洒落本の概要：全 17 作品．作品詳細は 5.1 節を参照．

地の文/会話文	地域	総文数	総短単位数	総文字数
地の文	江戸	4,255	29,983	50,348
	上方	2,384	19,525	32,451
	計	6,639	49,508	82,799
会話文	江戸	4,282	47,738	79,985
	上方	2,523	31,109	51,929
	計	6,805	78,847	131,914
	計	13,444	128,355	214,713

る近松門左衛門の世話物浄瑠璃である．この中で最も整備が進んでいるのは，『洒落本大成』(洒落本大成編集委員会 1978-88) を底本とした CHJ 洒落本であり，既に短単位アノテーション済みコーパスが 3 作品試験公開されているが(3)，国語研の内部的には，これ以上の資料についてアノテーションが進められている．また CHJ 近松に関してもアノテーションが進みつつあり，洒落本用の解析辞書 (市村太郎ほか 2016) を使った自動解析結果を順次人手で修正する作業を行なっている．

しかしながら文献 (市村太郎ほか 2016) にも取り上げられている通り，洒落本用の解析辞書を使った解析結果の精度は他の時代の辞書に比べて低い．文献 (小木曽智信ほか 2013) によると，整備が進んでいる中古和文系資料，および近代文語論説文では発音形認定のレベルで 96〜97 の F1 値を達成しているが，最新の文献 (市村太郎ほか 2016) の報告でも洒落本については，未だに語彙素認定のレベルで F1 値 90 台に止まっている(4)．文献 (小木曽智信ほか 2013) においては，語彙素認定で F1 値 95 を達成するためには最低でも 5 万短単位が必要と言われている．しかし表 1 を見ればわかる通り，CHJ 洒落本はすでにその要件を満たしている．

この問題について，文献 (市村太郎 2014) では以下の 2 つの原因を上げている．

問題 I. 表記の多様性　近世の資料の中には，一貫しない仮名遣いや送り仮名，片仮名平仮名の混ぜ書き，踊字の使用，濁点無表記 (岡照晃ほか 2013) といった，多様な表層形が出現し，辞書に網羅的に登録することが難しい(5)．

e.g., 動詞「言う」+ 接続助詞「て」で構成される表層形のバリエーションの一部：
「いつ｜て」「いゝ｜て」「言つ｜て」「いひ｜て」「いう｜て」「言ふ｜て」「云ひ｜て」「言｜て」「云｜て」「ゆう｜て」「いツて」(6)

問題 II. 地の文-会話文間の文体・文法体系の混在　CHJ の活用型が大きく文語活用と口語活用に分かれているのに対し，原文一致以前の資料である洒落本では，地の文は文語，会話文は口語で書かれている．そのため地の文と会話文とで活用型のアノテーションが別となり，地の文と会話文を辞書（=CRF のモデル）の学習・解析時に混用していると，活用型の認定でエラーが多発する．

これらに対し，問題 I については，既に文献 (市村太郎 2014) で，アノテーションレベルでの

(3) http://pj.ninjal.ac.jp/corpus_center/chj/edo.html
(4) いずれも「未知語なし」という条件の下での評価．
(5) 網羅したとしても，辞書サイズが肥大化し，かつ解析上の曖昧性も大きくなる．
(6) 「｜」は短単位の境界を表す．

対応がある程度図られており，文献 (岡照晃ほか 2014) では，その人手作業を自動化する試みが提案されている．問題 II については文献 (市村太郎ほか 2016) で，地の文と会話文とに分けて辞書を学習・解析する手法が提案されている．

しかしながら先に述べた通り，最新の文献 (市村太郎ほか 2016) でも洒落本解析用辞書の解析結果の精度は未だに他の時代の辞書に追い付いていない．問題 I が完全には解決されていないことも大きな原因ではあるが，文献 (市村太郎 2014) で言及されていないさらなる問題として，CHJ 洒落本の中に上方（関西）と江戸（関東），2 つの地域の資料が混在していることがある．

問題 III. 江戸-上方間の文体・文法体系の混在　例えば，先に上げた動詞「言う」+ 接続助詞「て」の表層形の一つである「言｜て」は，江戸では「イッ｜テ」と発音形がアノテーションされるが，上方では方言の違いから「ユー｜テ」とアノテーションされる．しかし江戸の資料と上方の資料が混在する状態で辞書の学習を行なった場合（図 1 の辞書 ii) ），「言｜て｜おく」(7)の発音形の解析結果は，上方の資料でも「イッ｜テ｜オク」となる．これは表 1 を見ればわかる通り，上方の資料よりも江戸の資料の方がデータとして多いため，解析時に曖昧な箇所が江戸側に倒れてしまうからである．もし逆に，上方の資料の方が多かった場合は，この解析結果は江戸・上方両方の資料で「ユー｜テ｜オク」となるだろう．また語彙素認定に関わる問題例として，助動詞「んす」がある．「んす」は語彙素「んす」の表層形として，洒落本解析用辞書にエントリされているだけでなく，語彙素「ます」の表層形としてもエントリされている．そして後者は主として江戸の資料にしか現れないという特徴がある．

このように，CHJ 洒落本の中に混在している文体・文法体系は地の文-会話文の別だけでなく，江戸-上方という方言の違いもある．この問題を解決するための方法として，地の文-会話文の問題を解決したときと同じように，辞書の学習用コーパスを江戸と上方で更に分割するという対応が考えられる．ただし，そもそも活用型から異なるような地の文-会話文の差に対して，江戸-上方の文体・文法体系の違いは著しく大きいわけではない．そのため学習用コーパスの縮小により，現状よりも解析結果の精度が悪くなる可能性がある．

3. 提案手法: MeCab の追加学習機能を用いた短単位解析用辞書の段階的特殊化

3.1 追加学習を用いた洒落本用解析辞書からの洒落本（江戸）・洒落本（上方）解析用辞書の作成

前節の問題 III に対し，本論文では，まず文献 (市村太郎ほか 2016) と同様に，地の文-会話文の別だけで 2 つの辞書を作る．そして次にそれぞれの辞書のモデルを江戸-上方の別で絞り込んだ学習用コーパスで MeCab の追加学習（**Regularized Adaptation**(8)）を実行する方針を採る．これにより，学習用コーパスが江戸-上方に分けて少量化される問題を回避しつつ，それぞれに適応した辞書が実現できる．

(8) 初期のモデルパラメータをできるだけ変更せずに，新しい学習データにできるだけ適応するような新しいモデルを学習する手法．MeCab の-M オプション（初期モデルの指定）で使用可能．詳細は文献 (Imamura 2013) を参照．

表 2　本稿で使用した CHJ 近松の概要.

	文数	総短単位数	総文字数
地の文	729	11,055	17,991
会話文	1,317	18,520	29,404
計	2,046	29,575	47,395

上記の手法により，次の 6 つの解析用辞書が作成される．各辞書の番号 i)〜vi) は，図 1 および，実験結果の表と対応している．

i): CHJ 洒落本（江戸 + 上方）の地の文から学習された辞書
ii): i) の辞書に CHJ 洒落本（江戸）の地の文だけ使って追加学習した辞書
iii): i) の辞書に CHJ 洒落本（上方）の地の文だけ使って追加学習した辞書
iv): CHJ 洒落本（江戸 + 上方）の会話文から学習された辞書
v): iv) の辞書に CHJ 洒落本（江戸）の会話文だけ使って追加学習した辞書
vi): iv) の辞書に CHJ 洒落本（上方）の会話文だけ使って追加学習した辞書

4. 追加学習を用いた洒落本（上方）解析用辞書からの近世前期の上方資料『近松門左衛門 世話物浄瑠璃』用解析辞書の作成

『近松門左衛門 世話物浄瑠璃』は近世前期の上方資料であり，CHJ 洒落本全体と比較した場合，まず近世前期と後期で時代に差があるほか，資料自体が上方に限定されているという特徴がある．また洒落本と異なり，地の文と会話文の文体差が小さいことも特徴の一つである．ただし CHJ 近松では，地の文と会話文の認定をそれぞれ実施し，その上で活用型に文語-口語の別でアノテーションを施している．表 2 に示す通り，CHJ 洒落本よりも少ないながら，CHJ 近松も既に形態論情報アノテーションが完了している部分が存在する．そこでこのアノテーション作業をさらに効率化するため，前節の洒落本（上方）で追加学習した辞書（辞書 iii)，vi)）に，アノテーション済みの CHJ 近松を追加学習することで CHJ 近松に特化した辞書を作成する．

よって，前節の辞書 i)〜vi) に加えて，さらに以下の 2 つの解析用辞書が作成される．各辞書の番号 vii), viii) は，図 1 および，実験結果の表と対応している．

vii): 前節 iii) の辞書に CHJ 近松の地の文だけ使って追加学習した辞書
viii): 前節 vi) の辞書に CHJ 近松の会話文だけ使って追加学習した辞書

5. 洒落本解析用辞書から近松解析用辞書までの段階的特殊化の性能評価実験

5.1 実験設定

本実験では，前節までに述べた以下の 8 つの辞書（再掲）を図 1 のように追加学習によって段階的に構築していき，各評価用コーパスでの性能を評価する．各辞書の番号 i)〜viii) は，図 1 および，実験結果の表と対応している．

i): CHJ 洒落本（江戸 + 上方）の地の文から学習された辞書
ii): i) の辞書に CHJ 洒落本（江戸）の地の文だけ使って追加学習した辞書
iii): i) の辞書に CHJ 洒落本（上方）の地の文だけ使って追加学習した辞書

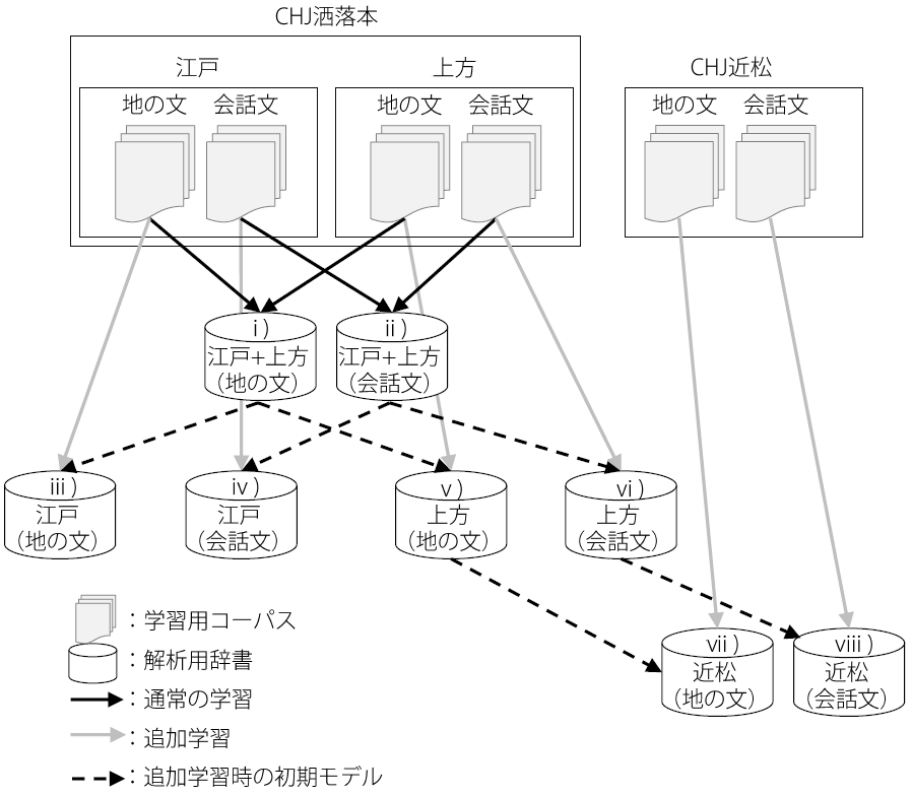

図 1　MeCab の追加学習機能を利用した解析用辞書の段階的特殊化.

iv): CHJ 洒落本（江戸 + 上方）の会話文から学習された辞書
v): iv) の辞書に CHJ 洒落本（江戸）の会話文だけ使って追加学習した辞書
vi): iv) の辞書に CHJ 洒落本（上方）の会話文だけ使って追加学習した辞書
vii): iii) の辞書に CHJ 近松の地の文だけ使って追加学習した辞書
viii): vi) の辞書に CHJ 近松の会話文だけ使って追加学習した辞書

表 3 と，表 4 にそれぞれ学習・評価用コーパスの内訳を示す．また整備中の CHJ 洒落本より使用する資料は以下の 17 作品である．

- 江戸
  - 郭中奇譚
  - 南閨雑話
  - 甲駅新話
  - 当世左様候
  - 花街鑑
  - 花街寿々女
  - 総籬
  - 仕懸文庫
- 上方
  - 原柳巷花語
  - 無論里問答

表 3　学習用コーパスの内訳.

地の文/会話文	学習用コーパス	総文数	総短単位数	総文字数
地の文	江戸（CHJ 洒落本）	3,830	27,016	45,412
	上方（CHJ 洒落本）	2,146	17,373	28,811
	江戸 + 上方（CHJ 洒落本）	5,976	44,389	74,223
	CHJ 近松	657	9,994	16,274
会話文	江戸（CHJ 洒落本）	3,854	43,312	72,501
	上方（CHJ 洒落本）	2,271	27,799	46,462
	江戸 + 上方（CHJ 洒落本）	6,125	71,111	118,963
	CHJ 近松	1,186	16,632	26,404

表 4　評価用コーパスの内訳.

地の文/会話文	評価用コーパス	総文数	総短単位数	総文字数
地の文	江戸（CHJ 洒落本）	425	2,967	4,936
	上方（CHJ 洒落本）	238	2,157	3,640
	江戸 + 上方（CHJ 洒落本）	663	5,119	8,576
	CHJ 近松	72	1,061	1,717
会話文	江戸（CHJ 洒落本）	428	4,426	7,484
	上方（CHJ 洒落本）	252	3,310	5,467
	江戸 + 上方（CHJ 洒落本）	680	7,736	12,951
	CHJ 近松	131	1,888	3,000

- 虚辞先生穴賢
- 阿闍陀鏡
- 昇平楽
- 誰か面影
- 興斗月
- 風流裸人形
- 箱まくら

解析用辞書のエントリはすべての辞書で共通とし，文献 (鴻野知暁ほか 2014) の時代情報を使った絞り込みによって UniDic データベース (小木曽智信ほか 2014) から取り出した 1,442,977 の表層形（辞書のキー）がエントリされている．このため次節に記載の実験結果は文献 (市村太郎ほか 2016) と同じく，未知語なしの条件下で実施したものとなっている．

MeCab の学習時の引数は，並列化オプションを除きすべてデフォルトのまま使用した．また CRF の正規化項のハイパーパラメータ C ($=\sigma^2$) も，全学習において共通にデフォルトの 1.0 に設定した．MeCab の学習時に使用する設定ファイルは文献 (市村太郎ほか 2016) と同じものを使用している．

評価は文献 (小木曽智信ほか 2013, 小木曽智信ほか 2014) と同じく，境界認定，品詞認定，語彙素認定，発音認定の 4 段階で F1 値を評価する．日本語の自動形態素解析における F1 値の計算方法は文献 (Kudo et al. 2004) を参照．境界認定は，文中での開始位置と終了位置が両方正しく認定できた（正解データと一致した）短単位数を評価している．品詞認定は，境界認定をパスした短単位の内，品詞大分類，中分類，小分類，細分類，活用型，活用形がすべて正しく

表 5　各評価用コーパス「地の文」における各辞書の短単位自動解析結果の精度（F1 値）.

評価用コーパス（地の文）	辞書	学習用コーパス	境界認定	品詞認定	語彙素認定	発音認定
江戸 + 上方	i)	江戸（地の文）+ 上方（地の文）	**97.19**	**91.93**	**90.11**	**89.43**
	ii)	追加学習：→江戸（地の文）	96.88	91.38	89.45	88.70
	iii)	追加学習：→上方（地の文）	97.09	91.45	89.58	88.83
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	96.88	89.70	87.99	87.15
	v)	江戸（会話文）+ 上方（会話文）	95.85	77.18	75.00	74.35
	vi)	追加学習：→江戸（会話文）	95.51	76.50	74.43	73.73
	vii)	追加学習：→上方（会話文）	95.91	74.99	72.39	71.73
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	95.81	74.76	73.02	72.42
江戸	i)	江戸（地の文）+ 上方（地の文）	**97.23**	**92.00**	**90.48**	**89.91**
	ii)	追加学習：→江戸（地の文）	97.10	91.94	90.42	89.75
	iii)	追加学習：→上方（地の文）	96.90	90.99	89.27	88.67
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	96.91	89.75	88.30	87.59
	v)	江戸（会話文）+ 上方（会話文）	95.48	77.17	74.95	74.38
	vi)	追加学習：→江戸（会話文）	95.28	76.79	74.60	73.96
	vii)	追加学習：→上方（会話文）	95.60	74.35	71.43	70.79
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	95.50	73.97	72.28	71.77
上方	i)	江戸（地の文）+ 上方（地の文）	97.14	91.83	89.60	88.76
	ii)	追加学習：→江戸（地の文）	96.58	90.62	88.10	87.26
	iii)	追加学習：→上方（地の文）	**97.35**	**92.09**	**90.00**	**89.06**
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	96.83	89.62	87.57	86.55
	v)	江戸（会話文）+ 上方（会話文）	96.37	77.20	75.06	74.31
	vi)	追加学習：→江戸（会話文）	95.83	76.10	74.19	73.40
	vii)	追加学習：→上方（会話文）	96.35	75.87	73.73	73.03
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	96.23	75.86	74.05	73.30
CHJ 近松	i)	江戸（地の文）+ 上方（地の文）	96.59	89.59	86.85	85.34
	ii)	追加学習：→江戸（地の文）	96.45	88.97	85.85	84.52
	iii)	追加学習：→上方（地の文）	96.03	89.12	86.09	84.58
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	96.60	**91.30**	**89.60**	**88.28**
	v)	江戸（会話文）+ 上方（会話文）	**97.25**	76.99	74.24	72.92
	vi)	追加学習：→江戸（会話文）	97.02	76.31	73.29	71.87
	vii)	追加学習：→上方（会話文）	96.60	76.86	74.03	72.52
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	97.21	80.28	78.39	77.16

認定できた短単位数を評価している．語彙素認定は，品詞認定をパスした短単位の内，語彙素読み，語彙素の両方が正しく認定できた短単位数を評価している．発音認定は，語彙素認定をパスした短単位の内，発音形出現形が正しく認定できた短単位数を評価している．評価には形態素解析器性能評価ツール『MevAL』[9]を使用した．

5.2　形態素解析性能評価実験の結果

辞書 i)～viii) を使った各評価用コーパス地の文の解析結果の精度の比較を表 5，会話文の解析結果の精度の比較を表 6 にそれぞれ示す．表 5 を見ると，地の文の境界認定に関しては，地の文で学習した辞書と会話文で学習した辞書の間で精度に大きな差は見られない．一方で，表 6 を見ると，洒落本の会話文の境界認定では，地の文で評価した場合よりも両辞書間での精

(9) https://teru-oka-1933.github.io/meval/

表 6　各評価用コーパス「会話文」における各辞書の短単位自動解析結果の精度（F1 値）.

評価用コーパス（会話文）	辞書	学習用コーパス	境界認定	品詞認定	語彙素認定	発音認定
江戸 + 上方	i)	江戸（地の文）+ 上方（地の文）	92.65	70.90	69.35	68.78
	ii)	追加学習：→江戸（地の文）	91.72	69.65	68.05	67.45
	iii)	追加学習：→上方（地の文）	92.56	70.27	68.89	68.34
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	92.73	70.64	69.18	68.61
	v)	江戸（会話文）+ 上方（会話文）	**97.64**	**92.67**	**91.69**	**91.08**
	vi)	追加学習：→江戸（会話文）	97.33	91.93	90.93	90.30
	vii)	追加学習：→上方（会話文）	97.36	90.98	89.92	89.29
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	96.64	88.48	87.16	86.52
江戸	i)	江戸（地の文）+ 上方（地の文）	92.31	70.28	68.87	68.37
	ii)	追加学習：→江戸（地の文）	91.78	69.66	68.21	67.71
	iii)	追加学習：→上方（地の文）	92.15	69.37	68.10	67.60
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	92.53	70.03	68.76	68.31
	v)	江戸（会話文）+ 上方（会話文）	**97.75**	**92.95**	**92.25**	**91.73**
	vi)	追加学習：→江戸（会話文）	97.41	92.94	92.17	91.70
	vii)	追加学習：→上方（会話文）	97.34	89.95	89.07	88.53
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	96.51	87.63	86.57	86.09
上方	i)	江戸（地の文）+ 上方（地の文）	93.10	71.73	70.00	69.33
	ii)	追加学習：→江戸（地の文）	91.65	69.64	67.84	67.11
	iii)	追加学習：→上方（地の文）	93.10	71.48	69.96	69.32
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	92.99	71.47	69.74	69.01
	v)	江戸（会話文）+ 上方（会話文）	**97.48**	92.30	90.94	90.22
	vi)	追加学習：→江戸（会話文）	97.22	90.58	89.28	88.44
	vii)	追加学習：→上方（会話文）	97.39	**92.35**	**91.06**	**90.30**
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	96.82	89.61	87.95	87.08
CHJ 近松	i)	江戸（地の文）+ 上方（地の文）	95.12	76.06	72.51	71.39
	ii)	追加学習：→江戸（地の文）	94.42	75.08	71.63	70.67
	iii)	追加学習：→上方（地の文）	94.57	75.60	72.16	70.99
	iv)	追加学習：→上方（地の文）→ CHJ 近松（地の文）	95.68	77.50	75.06	73.95
	v)	江戸（会話文）+ 上方（会話文）	95.13	87.56	85.07	83.32
	vi)	追加学習：→江戸（会話文）	95.05	86.95	84.36	82.61
	vii)	追加学習：→上方（会話文）	95.55	87.45	84.75	83.22
	viii)	追加学習：→上方（会話文）→ CHJ 近松（会話文）	**95.81**	**89.98**	**88.82**	**88.86**

度差が大きいことが分かる．これは，地の文が文語体で比較的漢語比率が大きく，解析に曖昧性が低いのに対し，会話文は仮名文字率が高く，分割の曖昧性が高くなったためと考えられる．しかし CHJ 近松での境界認定の評価では，地の文で学習した辞書と会話文で学習した辞書の間で洒落本ほど大きな差は見られなかった．これは近松の文体が文語・口語の区別の薄い関係だと思われる．

しかしながら品詞認定になると，洒落本と CHJ 近松の両方において，真逆の文体で学習した辞書の精度が同一文体で学習した辞書よりも格段に悪くなった．これは CHJ のアノテーション方針で，地の文には文語の活用型，会話文には口語の活用型をそれぞれアノテーションするという仕様が原因である．地の文のみで学習した辞書であれば文語活用，会話文で学習した辞書であれば口語活用が優先して自動付与するようになるが，評価用辞書が逆転すれば，地の文に口語活用，会話文に文語活用を適用してしまう．地の文-会話文の区別が薄い CHJ 近松でも，

会話文を同定し，地の文と分けて活用型を付与しているため，境界認定とは異なって，洒落本と同様の結果となっている．

次に追加学習の効果を見ると，評価用コーパス:上方では，上方で追加学習した場合（辞書 iii), vii) ）に，会話文の境界認定を除いて，江戸 + 上方で学習した場合（辞書 i), v) ）よりも精度の向上が見られた．これに対し，評価用コーパス:江戸になると，江戸 + 上方で学習した場合（辞書 i), v) ）の方が，江戸を追加学習した場合（辞書 ii), vi) ）よりも精度が高くなった．これはそもそも，今回の学習用コーパスのサイズが江戸の方に偏っていたことに加え，江戸の洒落本においても上方を真似して発話している箇所がいくつも存在しているためだと考えられる．また江戸 + 上方で学習した辞書（辞書 i), v) ）と江戸を追加学習した辞書（辞書 ii), vi) ）を使って文字列「言ておく」を解析した結果，いずれも発音形は「イッ｜テ｜オク」と解析されたが，上方を追加学習した辞書（辞書 iii), vii) ）では発音形は「ユー｜テ｜オク」と解析できることを確認した．

最後に評価用コーパス:CHJ 近松の場合，上方からさらに CHJ 近松で追加学習した結果が最も高い精度となったが，地の文の境界認定の評価でのみ，会話文の江戸 + 上方で学習した場合（辞書 v) ）の精度が一番高くなった．これは前述の通り，近松では地の文-会話文の差が薄く，追加学習するよりも，量の多い会話文だけで学習したほうが優位となったためだと思われる．ただし，活用型の文語-口語アノテーションの差から，品詞認定以降の評価では，地の文で追加学習した結果の方が高い性能となった．

6. おわりに

本論文では，近世前期の上方資料である近松門左衛門の世話物浄瑠璃への短単位形態論情報アノテーションの効率化を目的に，洒落本解析用辞書から段階的な追加学習により，近松解析用辞書を構築する方法について述べた．CHJ 洒落本コーパスには，江戸と上方，2 つの地域の資料が混在しており，それらを同時に辞書の学習に使用することについての問題点を指摘し，江戸の資料よりもサイズの小さい上方資料では，実際に提案手法によって解析結果の精度向上を確認した．また上方の洒落本資料から，さらに少量の CHJ 近松を追加学習して作成した近松解析用辞書を使うことで，従来の洒落本用解析辞書よりも高い精度で CHJ 近松の解析が行えることが分かった．現在，この近松用短単位自動解析用辞書を使ったアノテーション作業が進行中である．

近松資料の特徴として，地の文と会話文の文体差が薄いことがあり，実験結果の境界認定の精度評価でもそれは確認できた．しかし CHJ 近松ではあえて，地の文と会話文を別に認定し，文語活用-口語活用の別で活用型をアノテーションしている．そのため境界認定においては，地の文で学習した辞書よりも高い精度を出した会話文用の辞書でも，品詞認定より先の評価では地の文解析用辞書より低い精度となった．このことから，近松の短単位解析用辞書の解析結果の精度をさらに向上させる方法として，地の文用の辞書を学習する際でも，会話文を，活用だけを「*」に変えて使用し，会話文用の辞書でもそれと逆の処理を実行する，という試みが今後の展開として考えられる．

謝　辞

本研究は，国立国語研究所共同研究「通時コーパスの構築と日本語史研究の新展開」の研究成果を報告したものである．

文　献

[Imamura 2013] Kenji Imamura (2013). “Case Study of Model Adaptation: Transfer Learning and Online Learning.” *Proceedings of IJCNLP-2013 (the 6th International Joint Conference on Natural Language Processing)*, pp. 1292–1298.

[Kudo et al. 2004] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto (2004). “Applying Conditional Random Fields to Japanese Morphological Analysis.” *Proceedings of EMNLP-2004 (the 2004 Conference on Empirical Methods in Natural Language Processing)*, pp. 230–237.

[Lafferty et al. 2001] John Lafferty, Andrew McCallum and Fernando Pereira (2001). “Conditional random fields: Probabilistic models for segmenting and labeling sequence data.” *Proceedings of ICML-2001 (the 18th International Conference on Machine Learning)*, pp. 282–289.

[市村太郎 2014] 市村太郎 (2014).「近世口語資料のコーパス化–狂言・洒落本のコーパス化の過程と課題–」日本語学 11 月臨時増刊号 日本語史研究と歴史コーパス 33:14, pp. 96–109.

[市村太郎ほか 2016] 市村太郎・小木曽智信 (2016).「文書構造を利用した近世期洒落本の形態素解析」言語処理学会 第 22 回年次大会 発表論文集，pp. 107–110.

[小木曽智信ほか 2013] 小木曽智信・小町守・松本裕治 (2013).「歴史的資料を対象とした形態素解析」自然言語処理，20:5, pp. 727–748.

[小木曽智信ほか 2014] 小木曽智信・中村壮範 (2014).「『現代日本語書き言葉均衡コーパス』形態論情報アノテーションシステムの設計・実装・運用」自然言語処理，21:2, pp. 301-332.

[岡照晃ほか 2013] 岡照晃・小町守・小木曽智信・松本裕治 (2013).「統計的機械学習を用いた歴史的資料への濁点付与の自動化」情報処理学会論文誌，54:4, pp. 1641–1654.

[岡照晃ほか 2014] 岡 照晃・松本裕治 (2014).「形態素解析との同時解析による歴史的資料の自動表記整理」情報処理学会第 216 回自然言語処理研究会 情報処理学会研究報告, 2014-NL216:8, pp. 1–20.

[鴻野知暁ほか 2014] 鴻野知暁・小木曽智信 (2014).「見出し語の時代情報を付与した電子化辞書の構築」言語処理学会 第 20 回 年次大会 発表論文集，pp. 209–212.

[近藤泰弘 2012] 近藤泰弘 (2012).「日本語通時コーパスの設計」NINJAL「通時コーパス」プロジェクト・Oxford VSARPS プロジェクト合同シンポジウム「通時コーパスと日本語史研究」予稿集，pp.1–10.

[近藤泰弘 2015] 近藤泰弘 (2015).「『日本語歴史コーパス』と日本語史研究」コーパスと日本語史研究，ひつじ研究叢書＜言語編＞，第 127 巻， ひつじ書房，pp.1–16.

[洒落本大成編集委員会 1978-88] 洒落本大成編集委員会 編 (1978-1988).「洒落本大成」中央公論社.

[伝康晴ほか 2007] 伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元 清貴・小磯 花絵 (2007).「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」『日本語科学』，22 号，pp.101–123.